DOCUMENT RESUME

ED 055 792 SE 011 353

AUTHOR	Bauman, Daniel J.; And Others
TITLE	A Guide to the Instruments Developed for Evaluation of In-Service Institutes.
PUB DATE	Apr 70
NOTE	9p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Evaluation; *Evaluation Techniques; *Inservice Teacher Education; *Measurement Instruments; Program Evaluation; *Science Institutes
IDENTIFIERS	Earth Science Curriculum Project

ABSTRACT

The purpose of this paper is to provide summary data relative to the instruments developed for analysis of in-service institutes. This paper initiates a slightly different form of offering instruments for use by others. Rather than presenting a finished instrument with only summary data provided and all of the subjective decisions made, the reader is given the opportunity to make his own decisions and, using basic data which must be requested from the publisher, to develop his own instruments. The developed instruments may also be requested, along with summary data such as factor loadings, reliability, standard error of measurement and indications of needs. The development and application of the instruments is described, but content of the instruments is not provided in this paper. (Author/PR)

A GUIDE TO THE INSTRUMENTS DEVELOPED FOR EVALUATION OF IN-SERVICE INSTITUTES

by Daniel J. Bauman, Larry A. Irwin and John F. Thompson

The purpose of this paper is to provide summary data relative to the instruments developed for analysis of in-service institutes. With this paper, we would like to initiate a slightly different form of offering instruments for use by others. Rather than presenting a finished instrument with merely summary data provided and all of the subjective decisions made, we would like to offer the basic data, allow you to make your own decisions and, using this basic data, develop your own instruments. At the same time, we offer our revised instruments which are in the mode typically offered--that is, we will provide instruments in the form that we intend to use along with summary data such as factor loadings, reliability, standard error of measurement and indications of what we feel we are measuring. This may or may not agree with your particular needs. If you are engaging in a major effort where the expense involved is justified, we would like to recommend that you come to Boulder and spend a day or two examining the basic data that we have on each item, including the items that we have rejected. We encourage you to make your own set of decisions as to which items should be used and which items should not, and how the items should be grouped.

We are very much aware of the possibility of multiple interpretations of the empirical data. We are also aware of how the subjective judgments made in establishing which items are retained and which items are not can affect the final instrument developed. In addition to the empirical data, preference for one phrasing over another phrasing enters into a decision to retain or reject an item. When we make those decisions here, we are making them with reference to a particular group that we plan to test. Personal bias enters into every

decision made. By offering you the raw data, we are offering you the chance to make your own decisions based on your own biases. If your situation does not warrant your spending money on the examination of the basic data, we believe that we have made a reasonable judgment and have a very respectable set of instruments that can be used as is. We suggest that persons desiring to use them as is write and ask for the revised instruments. These are the instruments that are being used during the 1969-70 evaluation.

Each year we intend to add items and perhaps questionnaires or instruments. As we add them, these other questionnaires or instruments will also be available along with the empirical data that we gather on them. Most of these instruments are in the affective domain and, therefore, are subject to interpreptation and also subject to changes due to history and other variables. As the years pass events happen that change the response patterns to these instruments. Therefore, especially with the affective domain, keep in mind that we intend a continuous process of updating and revision. This will be an evolutionary process.

The method used to develop these instruments was first to construct a data matrix. We obtained consensus on 32 questions about institutes which would be of interest to us. A few examples of the sort of question we chose are the following: "Has teacher behavior changed?" "Description of teachers' sensitivity to student needs." "Teachers' ability to teach earth science courses."

Following the selection of questions, we wrote as many items as we could that might be relevant to each question. In writing the items, we specified who was to respond to them. These item-writing sessions were a series of brainstorming sessions where all conceivable items were listed. These sessions were followed by sessions to reduce and refine those items. Appropriateness to the recipient of the questionnaire, probable measurement yield, and relevance were the principal criteria used.

The next step was grouping the items by respondents so that all items to be answered by the institute participant were grouped together, all items to be responded to by the administrator were grouped together, etc. Each group now will be referred to as a questionnaire. Thus we have the participant questionnaire, the administrator questionnaire, the student questionnaire, the supervisor questionnaire, teacher questionnaire, institute evaluation inventory, and earth science attitude inventory. Any individual instrument was not homogeneous but consisted of several different questions being answered. Thus, analysis by questionnaire would not be a logical activity.

Analyses were by question from the original data matrix. For the remainder of this paper, the group of items referring to one question will be referred to as a subscale. I would like to emphasize that the analysis is by subscale. No analysis was run on the teacher questionnarie as a questionnaire, nor was there an analysis run on the director questionnaire. Analysis was by subscale. Our unit in most of the analyses was the participant. The responses of the students were grouped by participant and analyzed by participant. Twenty-seven of the subscales were such that we could use the participants as the unit. On the 27 participant subscales, we had a total of 379 items. No analysis was run on the 379 items as a group. Five subscales were appropriate to the institute as the unit and were analyzed in a separate group. See Table 1 for summary statistics on these 32 subscales. There were 96 items with the institute as the unit of analysis.

Our sample population of participants was 750. We randomly sampled one-third of each institute and obtained complete responses on 214 participants. Students of these 214 participants were sampled, three per science class. Thus, a teacher with five science classes would be represented by 15 student response sets. We obtained 3112 student response sets. Since we had only 26 institutes, the

institute directors and institute instructors were not reduced by sampling. We used all of the institute directors and all of the institute instructors in the questionnaires directed at them.

All item analyses were run on the computer program Fortap described by Frank Baker in Educational and Psychological Measurement, spring 1969. We obtained the program from Dr. Baker and made modifications to input and output sections. We did not alter the basic computation package. After making our modifications, we did confirm the computational accuracy of the program by running the BMD analysis of variance program to confirm the Hoyt analysis of the variance reliability estimate. We also ran the BMD program on Pearson product correlation to compare with the point bi-serial correlation in the Fortap Program. The Hoyt reliabilities checked identically. The correlations were within the expected agreement range. From the Fortap program we were able to apply the Hoyt estimate of reliability, the bi-serial correlation between item response and total score, the mean and standard deviations and the number responding with each option to each item for each subscale. On a number of the subscales we ran factor analyses using the BMD program for factor extraction employing squared-multiple correlation as estimate of communality.lity. We followed this factor extraction with the oblique rotation suggested by Harris-Kaiser as programed at the University of Wisconsin. The oblique program was also furnished to us by Frank Baker. We are reporting the pattern matrix from the oblique solution as our factor loadings. As a check on our factor solution, we selected what we felt was the most divergent commonly accepted factor analysis routine, the routine of Tryon and Bailey employing cluster analysis. This method employs quite different criteria for defining clusters. In a comparison of these two methods on one problem we found no difference with this method in the final interpreted results. There were small differences in the factor loadings. While there were difference in that Factor 2 in one

program came out as Factor 5 in the other program, the items defining factors were the same. Since there was no difference, we did not continue to run the Tryon and Bailey program on the other analyses. All of the problems then were run with squared-multiple correlations as communality estimates in the BMD principal component factor extraction program followed by the oblique solution of Harris-Kaiser. A sample of the factor loadings on one of the subscales is included as Table 2 of this report. We have included in this table a number of loadings that are too small to be interpretable. Only the smallest loadings, those less than .1, were omitted from the table. By being inclusive in this manner, you will be able to make some judgments regarding our interpretation of the analysis. This is the sort of table that we will provide unless you specify otherwise. We do have the full data matrix and the full structure matrix, as well as the original orthogonal factor extraction available.

One concern when we are combining information from different sources is that the different sources might introduce bias to the data. We need to indentify the different sources and the items coming from each source. If bias exists, the different sources would come out as different factors in the factor analysis. This would also show up to some extent in the item analysis in that the correlations of responses with total score would differ by groups according to the source of the items. By having information from different sources, we are also, in effect, running a validity check between sources. Data is provided regarding these two related questions. We feel satisfied with the results.

Since each of the subscales consisted of items directed toward a single question, we would expect to get one single large factor out of a factor analysis of a subscale and very weak secondary factors. This, indeed, proves to be the case. As an example, on one subscale the first eigenvalue was 13.85; the second eigenvalue was less than 2. That we had single strong factors indicated that

our items were measuring a single question rather than being divided among several questions. Where we do have meaningful second or third factors, they can be interpreted as measuring different aspects of the single, primary question. It is conceivable that some of the secondary factors are merely reflecting sentence or grammatical construction of the item.

Validity of instruments is always a primary concern. Brief mention was made of validity in the last section. Our aim has been and continues to be a validation by comparison with recognized instruments. One problem encountered here is the lack of good instruments to use in establishing validity. There are several instruments that we would like to have comparisons with, and we are attempting to perform those comparisons in our continuing program. Another means of establishing validity is to have an in-classroom observer making judgments during the institute. This is a very expensive procedure and is subject to considerable variation, depending upon the particular bias and subjectivity of the observer employed. In the meantime, we feel that we do have a reasonable cross-validation check in the comparisons of ratings by the participants, by the students, by the supervisors, by the director of the institute and by the instructors in the institutes. Certainly this method has advantages in some respects over the other methods of establishing validity. We feel that we do have good agreement between sources. If you can provide comparisons with other instruments or other means of validation, we would be very happy to exchange information with you. We are very much interested in obtaining validation information.

Another item of interest that came up during processing of this data pertained to omitted responses. Our attention was called to the problem when our initial run of item analyses showed extremely high reliabilities. All of these reliabilities were over .9 with a large number of them over .99. We had intended to follow the usual practice of omitting any response sets for which

appreciable numbers of responses were omitted and running the remainder on the usual program. It was suggested that we should not score those items that were omitted. Since omitting an item is equivalent to scoring it zero, this is not usually possible, especially on a scaled response item. For example, in a case where you have responses scaled from one to five, scoring the omits as zeros is a serious error and puts the response to the far end of the scale beyond the one. On a scale response where one is an extreme response to one end of the scale and five is an extreme response to the other end of the scale, the zero then puts us beyond the extreme. This is not a reasonable interpretation of an omitted response. It makes logical sense to put an omitted response in the middle of the scale rather than at an extreme position. In our scoring we have used this middle position as the weight for omits and obtained an item analysis reflecting the group responding with omits for each item.

Our present practice then is initially to score all omits at the middle of the scale for that particular item. We then examine the item analysis and follow this by weighting omits with the weight attributed to the group most closely matching the omits in their total scores. Let us look at the mechanics of this. We initially score omits with the middle of the scale response. In the item analysis we get a correlation, a bi-serial correlation, of omitting that item with total score. We also get bi-serial correlations between each of the other responses and total score. If the correlation between omits and total score is perhaps -.2 and the correlation of the total score with response No. 4 is approximately -.2, we would then assign weight to the omit that is the same as the weight assigned to response 4.

Since the quantity of data available is so large, please be specific in requesting materials. The financial burden of making a blanket distribution of all data collected would be too great to be assumed. For a summary of the

results from use of the instruments, see the article, "An Evaluation of 26 NSF-Funded ESCP In-Service Institutes," by John F. Thompson, available upon request from the Earth Science Teacher Preparation Project (ESTPP), P. O. Box 1559, Boulder, Colorado 80302. Copies of the instruments and further specific data can also be obtained from ESTPP.

DJB/od

4/29/70

TABLE 2

PATTERN MATRIX (A)

(SUBSCALE 9)

Variable												
1									.79			
2	.19	.13	.18	.17		.26	.28		.26			.26
3	.88											
4	.87											
5	.10				.80							.14
6	.15				.64		.12					.20
7	.02									.67		
8	.07	.15					.47					.15
9												.59
10		.76					.11	.10				
11		.54	.13				.11			.22		
12		.15						.54				
13			.13	.16	.18	.58		.18	.12	.21	.10	
14	.09	.11			.17	.65	.21		.23	.18		
15				.27		.69			.17			
16				.55					.11			
17		.17	.87					.13	.14			
18		.11	.83				.12		.16			
19		.28	.53					.32			.18	.14
20											.41	
Eigenvalues	4.911	2.520	1.590	.548	.320	.284	.203	.145	.119	.044	.021	.008